

# Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments

William R. SHADISH, M. H. CLARK, and Peter M. STEINER

A key justification for using nonrandomized experiments is that, with proper adjustment, their results can well approximate results from randomized experiments. This hypothesis has not been consistently supported by empirical studies; however, previous methods used to study this hypothesis have confounded assignment method with other study features. To avoid these confounding factors, this study randomly assigned participants to be in a randomized experiment or a nonrandomized experiment. In the randomized experiment, participants were randomly assigned to mathematics or vocabulary training; in the nonrandomized experiment, participants chose their training. The study held all other features of the experiment constant; it carefully measured pretest variables that might predict the condition that participants chose, and all participants were measured on vocabulary and mathematics outcomes. Ordinary linear regression reduced bias in the nonrandomized experiment by 84–94% using covariate-adjusted randomized results as the benchmark. Propensity score stratification, weighting, and covariance adjustment reduced bias by about 58–96%, depending on the outcome measure and adjustment method. Propensity score adjustment performed poorly when the scores were constructed from predictors of convenience (sex, age, marital status, and ethnicity) rather than from a broader set of predictors that might include these.

KEY WORDS: Nonrandomized experiment; Propensity score; Randomized experiment; Selection bias

## 1. INTRODUCTION

Randomized experiments can yield unbiased estimates of effect sizes. But randomized experiments are not always feasible, and other times ethical constraints preclude random assignment. Consequently, researchers often use nonrandomized experiments (Rosenbaum 2002; Shadish, Cook, and Campbell 2002) in which participants self-select into treatments or are selected nonrandomly to receive treatment by an administrator or service provider. Unfortunately, whatever feasibility or ethical benefits sometimes accrue to nonrandomized experiments, they yield effect estimates that either are demonstrably different from those from randomized experiments (Glazer, Levy, and Myers 2003) or are at best of unknown accuracy (Rosenbaum 2002). To explore the accuracy of estimates from nonrandomized experiments, previous research has compared randomized and nonrandomized experiments in one of three ways: computer simulations, single-study comparisons, or meta-analysis. All three approaches have weaknesses that the present study remedies. A fourth method that we discuss, the doubly randomized preference trial, works well in theory but in practice is plagued by problems of attrition and partial treatment implementation.

Computer simulations (e.g., Drake 1993) investigate these issues by generating precisely controlled but artificial data, varying key features that might affect results, such as the magnitude of the bias or the sample size. The high control and the large number of replications in these simulations yield very accurate results. But such simulations are quite artificial, for example, presuming that data are normally distributed or that outcome measures have no measurement error. Most importantly, simulations require the researcher to specify the selection model

for nonrandomized experiments; but in nonrandomized experiments, the problem is that the researcher does not know that model. So simulations can only approximate real-world selection bias problems, and they do so to an uncertain degree.

Two other methods provide more realistic contexts for studying selection bias (Shadish 2000). The single-study approach compares results from an existing randomized experiment with results obtained when a single nonrandomized control that is conveniently available is substituted for the original randomized control (or, alternatively, by comparing the randomized control with the nonrandomized control on the assumption that if the two control groups are equal, then the nonrandomized control can be substituted for the randomized control). This method gives the researcher access to raw data from individual participants, so that he or she can apply statistical adjustments to those data to improve the estimates. The results of such studies have been mixed, with some studies supporting the use of adjustments and others not doing so; for example, Heckman, Ichimura, and Todd (1997) randomly assigned applicants to a control group or to a job training program, and also collected data on a group of eligible nonparticipants who met the requirements for the training program but were not participating in it. They then compared the randomized treatment group both to the nonrandomized control group (the nonrandomized experiment) and to the randomized control group (the randomized experiment). The two experiments yielded different estimates when adjusted using econometric selection bias models. In comparison, more optimistic results were obtained in studies by Dehejia and Wahba (1999) and Hill, Reiter, and Zanutto (2004) using propensity score adjustments. Hill et al. (2004) also used multiple imputation to cope with the inevitable missing data that occur both before and after treatment in field experiments.

At first glance, studies like those of Dehejia and Wahba (1999), Heckman et al. (1997), and Hill et al. (2004) seem to provide a credible test of the effects of adjustments such as

William R. Shadish is Professor, Founding Faculty, University of California Merced, Merced, CA 95344 (E-mail: [wshadish@ucmerced.edu](mailto:wshadish@ucmerced.edu)). M. H. Clark is Assistant Professor of Psychology, Southern Illinois University, Carbondale, IL 62901 (E-mail: [mhclark@siu.edu](mailto:mhclark@siu.edu)). Peter M. Steiner is Assistant Professor, Institute for Advanced Studies, 1060 Vienna, Austria, and currently a Visiting Research Scholar at Northwestern University, Evanston, IL 60208 (E-mail: [steiner@ihs.ac.at](mailto:steiner@ihs.ac.at)). Shadish and Steiner were supported in part by grant 0620-520-W315 from the Institute for Educational Sciences, U.S. Department of Education.

propensity score analysis or selection bias modeling. However, these studies all share a key weakness that renders their results unclear—they confound assignment method with other study features. These confounds are problematic. Adjustments such as propensity score analysis are attempting to estimate what the effect would have been had the participants in a nonrandomized experiment instead been randomly assigned to the same conditions using the same measures at the same time and place. The latter counterfactual cannot be observed directly. As has been argued in causal inference in general (Rubin 1974; Holland 1986), the best approximation to this true counterfactual may be a group of participants whose assignment method (random or nonrandom) was itself randomly assigned to them, with all other features of the experiment held equal. This was not done by Dehejia and Wahba (1999), Heckman et al. (1997), or Hill et al. (2004), or in any other such studies. Rather, assignment mechanism (random or nonrandom) was varied nonrandomly in those studies and always was confounded with other differences between the random and nonrandom control groups. For example, compared with the randomized control group, the nonrandomized control group often was assessed at different sites or times, by different researchers, with different versions of the measure; and the groups may have had different rates of treatment crossover and missing outcome data. Even if these confounding factors were known, it would be impossible to adjust for some of them, because the single-study approach relies on just one instance of a randomized control and a nonrandomized control, so there is no variability in study-level confounding factors. Consequently, if research that uses the single-study approach finds that a selection bias adjustment to the nonrandomized experiment does (or does not) yield the same results as the randomized experiment, then we cannot know whether this is due to the adjustment method or to variability caused by these other confounding factors.

Meta-analysis offers a partial remedy to the problem of confounding factors by comparing many randomized and nonrandomized experiments on the same question to see whether they yield the same average effect size. Lipsey and Wilson (1993) used the simplest form of this approach, summarizing results from dozens of meta-analyses comparing randomized and nonrandomized experiments. The average over these comparisons was 0—nonrandomized experiments yielded the same effect size as randomized experiments on average—although in any given meta-analysis, the difference usually was not 0. But the validity of this overall average relies on the assumption that any variables that are confounded with assignment method are distributed randomly over meta-analyses. Data suggest that this is unlikely to be the case (e.g., Heinsman and Shadish 1996). In an attempt to lessen reliance on this assumption, other meta-analyses have coded such confounding factors and included them as covariates to get an adjusted difference between randomized and nonrandomized experiments (e.g., Heinsman and Shadish 1996; Shadish and Ragsdale 1996; Glazerman et al. 2003). These meta-analyses have yielded mixed results, with some concluding that the adjusted difference is near 0 (Heinsman and Shadish 1996) and others concluding that it is not (Glazerman et al. 2003).

Fundamentally, however, the meta-analytic approach suffers from the same flaw as the single-study approach, which is not

surprising because it is based on those single studies. Variables confounded with assignment mechanism are still unknown, and so the researcher cannot be sure that all relevant confounding covariates have been identified, measured well, and modeled properly. Moreover, the meta-analytic approach also cannot access primary raw data from each experiment, so it cannot test whether such adjustments as selection bias modeling or propensity score analysis improve estimates from nonrandomized experiments.

To address some of the problems with these methods, the present study explores the differences between randomized and nonrandomized experiments using a laboratory analog that randomly assigns participants to be in either randomized or nonrandomized experiments that otherwise are equal in all respects. This equating of experimental methods on conditions other than assignment method remedies the key weakness of both the single-study approach and the meta-analytic approach in which other variables can be systematically confounded with estimates of the effects of assignment method. The method also remedies the additional problem of the meta-analytic approach by producing data on individual participants, allowing the use of adjustments to reduce bias that are not available to the meta-analytic approach. Finally, the method examines naturally occurring selection biases in which the selection process is unknown, a more realistic test than in computer simulations.

The approach in the present study is related to a fourth method—the doubly randomized preference trial (DRPT) (Rücker 1989; Wennberg, Barry, Fowler, and Mulley 1993; Janevic et al. 2003; Long, Little, and Lin, in press)—although it differs in some important ways. First, some of the DRPT literature makes only hypothetical proposals about the possibility of implementing DRPTs (e.g., Wennberg et al. 1993) or is devoted only to developing a statistical model for assessing effects in DRPTs rather than to gathering experimental data with a DRPT (e.g., Rücker 1989). This is nontrivial, because the practical problems involved in executing DRPTs are formidable and, as we argue later, usually impede the ability of DRPTs to obtain a good test of the effects of adjustments, such as propensity score analysis. Second, none of the DRPT studies conducted to date has used the design to assess whether adjustments to observational studies like propensity score analysis can replicate results that would have been obtained had participants been randomized.

Third, and perhaps most importantly, because our method uses a brief laboratory analog treatment, it avoids problems of partial treatment implementation and of missing outcome data that occurred in the few past DRPTs that actually tried to gather data. This is crucial, because adjustments like propensity score analysis answer questions only about what would have happened to the participants in the nonrandomized experiment had they been randomly assigned to conditions. They do not adjust for partially implemented treatments or for missing outcome data, but any DRPT conducted in a field setting is almost certain to encounter both of these problems. For example, nearly two-thirds of those initially assigned to the conditions of Janevic et al. (2003) refused to accept their random assignment to the randomized or choice arms of the study, and all of them had missing outcome data. Although the differential rate of refusal (3%) to conditions was minimal (62% refusal to the choice arm

60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118

1 vs. 65% to the randomization arm), an additional 4% withdrew  
 2 from the choice arm after the pretest, making the differential  
 3 missing outcome data  $65\% - 58\% = 7\%$ . Moreover, such biases  
 4 might be differential in substantive nature across conditions if  
 5 those willing to accept no choice of condition (i.e., random assign-  
 6 ment) are different from those who are willing to participate  
 7 only if they can choose their conditions.

8 Janevic et al. (2003) also reported large and significant differ-  
 9 ences in treatment implementation rates between the random-  
 10 ization and choice arms of the study. The reanalysis of these  
 11 data by Long et al. (in press) used an intent-to-treat analysis  
 12 to estimate causal effects in the presence of such problems, but  
 13 that analysis cannot be done without additional assumptions be-  
 14 yond an adjustment for assignment method. Thus the resulting  
 15 comparison of the adjusted results from the randomized and  
 16 nonrandomized experiments of Janevic et al. (2003) is a joint  
 17 test of the effects of adjusting for assignment method, missing  
 18 outcomes, and partial treatment implementation. Our method  
 19 substantially avoids these two problems and thus allows testing  
 20 of the effects of adjustments for assignment method that are less  
 21 encumbered by extraneous concerns.

22 Our method has its own problems, however. What may be  
 23 gained in purity of the adjustment for assignment method using  
 24 the present method may be lost in questions about generaliza-  
 25 tion from the laboratory to the field, about the substantive im-  
 26 portance of the brief intervention, and about other issues that  
 27 we describe in more detail in Section 4. In addition, our method  
 28 represents only one kind of observational study, a prospective  
 29 nonrandomized experiment in which participants agree to be  
 30 recruited and to be randomized to randomization or choice con-  
 31 ditions. Those who agree to be recruited to such an experiment  
 32 may differ from those who self-select into a program of their  
 33 own accord, as might be more common in retrospective obser-  
 34 vational studies. Thus the present method is just one alterna-  
 35 tive with its own strengths and weaknesses compared with past  
 36 methods.

37 Nonetheless, the unique contribution of the present study is  
 38 the novel methodology for testing the accuracy of proposed sta-  
 39 tistical solutions to a critically important problem in statisti-  
 40 cal practice. Although at first glance there may be little moti-  
 41 vation for interest in a brief laboratory analog treatment, this  
 42 format is a key virtue, because it allows estimation of the ef-  
 43 fects of adjustments for nonrandom assignment unconfounded  
 44 with assumptions about missing outcome data, partial treatment  
 45 implementation, or other differences between the randomized  
 46 and nonrandomized experiments. Although one might imagine  
 47 a field experiment with similar virtues, such as a very brief med-  
 48 ical intervention that is fully implemented with an outcome that  
 49 is a matter of public record and in which participants readily  
 50 agree to be randomly assigned to whether or not they get a  
 51 choice of treatment, such a field experiment has yet to occur,  
 52 and its practical logistics would be formidable.

53 The rest of this article is organized as follows. Section 2 de-  
 54 scribes the method and its implementation. Section 3 presents  
 55 the results, with particular focus on propensity score adjust-  
 56 ments. Section 4 discusses the promise and the limitations of  
 57 this study and suggests ways of extending this methodology to  
 58 explore its generalizability.

## 2. METHODS

60 The study began with baseline tests that were later used to  
 61 predict treatment selection (Fig. 1). Then participants were ran-  
 62 domly assigned to be in a randomized experiment or a nonran-  
 63 domized experiment. Those assigned to the randomized exper-  
 64 iment were randomly assigned to mathematics or vocabulary  
 65 training. Those who were assigned to the nonrandomized exper-  
 66 iment chose which training they wanted and then attended the  
 67 same training sessions as those who were randomly assigned.  
 68 After training, all participants were assessed on both mathe-  
 69 matics and vocabulary outcomes. This design ensured that all  
 70 participants were treated identically in all respects except as-  
 71 signment method.

### 2.1 Participants

72 Volunteer undergraduate students from introductory psychol-  
 73 ogy classes at a large mid-southern public university were as-  
 74 signed randomly to be in a randomized ( $n = 235$ ) or a nonran-  
 75 domized ( $n = 210$ ) experiment, using month of birth for prac-  
 76 tical reasons. These sample sizes were not large, a limitation  
 77 if propensity scores are most effective with large samples. But  
 78 such sample sizes are common in applications of propensity  
 79 scores in field experimentation. Students received experimen-  
 80 tal credit that was either required or allowed for their classes,  
 81 and they chose to participate in this experiment from among  
 82 several available experiments. Of the 450 students who signed  
 83 up for the experiment, 445 completed the pretests, intervention,  
 84 and posttests. The remaining five participants dropped out af-  
 85 ter being assigned to conditions but during the transition from  
 86 pretest administration to training. Of these, three were ran-  
 87 domly assigned to the randomized experiment (two then ran-  
 88 domly assigned to mathematics and one to vocabulary), and two  
 89 were randomly assigned to the nonrandomized experiment (one  
 90 chose vocabulary, and one did not complete the choice form).  
 91 These five students were dropped from analyses because their  
 92 missing outcomes were only 1.1% of the data and because their  
 93 distribution was even over assignment to random versus non-  
 94 random experiments. These five were the only participants lost  
 95 to treatment or outcome measurement.

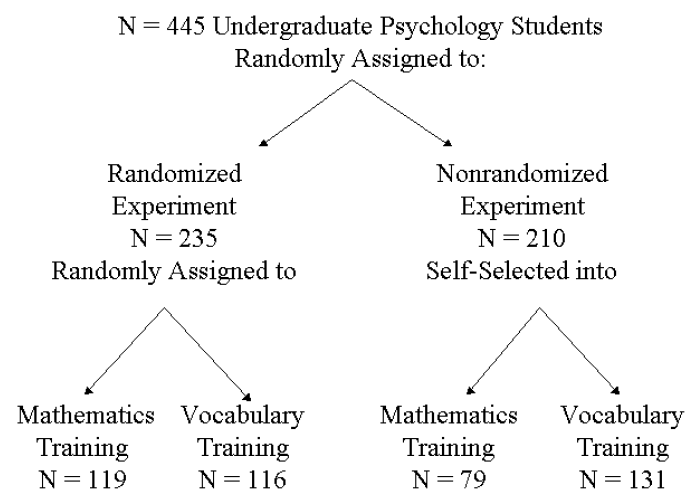


Figure 1. Overall design of the study.

## 2.2 Pretests

Written instructions and computer scored answer sheets were used for all of the following pretests:

- Demographics Questionnaire I, prepared by us, gathered data about participant age, education, marital status, major area of study, ACT and SAT scores, and grade point average (GPA) for college and high school.
- The Vocabulary Test II (Educational Testing Services 1962) measured vocabulary skills to predict selection into mathematics or vocabulary training.
- The Arithmetic Aptitude Test (Educational Testing Services 1993), administered with scratch paper, measured mathematics skills to predict selection into conditions.
- Demographics Questionnaire II, prepared by us based on an interview with a full-time staff member of the student educational advising center, assessed previous scholastic experiences in mathematics and vocabulary to predict selection into condition.
- The International Personality Item Pool test (Goldberg 1997) assessed five major domains of personality: extroversion, emotional stability, agreeableness, openness to experience, and conscientiousness.
- The Short Mathematics Anxiety Rating Scale (Faust, Ashcraft, and Fleck 1996) assessed stress induced by mathematics to predict selection into mathematics training.
- The Short Beck Depression Inventory (Beck and Beck 1972) assessed depression, given that a previous scale assessing depression in college students (Kleinmuntz 1960) predicted performance.

## 2.3 Treatments

A series of overhead transparencies presented interventions to teach either 50 advanced vocabulary terms or 5 algebraic concepts. The vocabulary transparencies each included a novel term, its phonetic spelling, and a sentence in which the word was used. The mathematics transparencies included five rules for transforming exponential equations and several examples in which those rules were applied to algebraic formulas. We compared two treatment conditions (rather than comparing treatment to no treatment) for two reasons: (a) Doing so created two effect estimates, one for the effects of vocabulary training on vocabulary outcome and one for the effects of mathematics training on mathematics outcome, and (b) a “no treatment” control might attract a disproportionate number of participants to select the least time-consuming session in the nonrandomized experiment. We chose to train participants in mathematics and vocabulary for three reasons. First, various kinds of mathematics and language skills are studied from elementary school through college, are often used in educational testing and are basic skills for many academic and career fields, so they are good analogs to topics sometimes studied in field experiments. Second, through experimental control over the difficulty of the vocabulary terms and algebraic concepts, we could anticipate that most participants would not be familiar with the material before the experiment and, correspondingly, anticipate that the experimental effect size would be meaningfully large. Third, college students differ greatly in their propensity to choose

mathematics training, reflecting a condition ripe for selection bias, thus making it easier to detect differences between randomized and self-selected conditions.

Training sessions were conducted by one of four white males, including three psychology graduate students and one undergraduate psychology major. Trainers were counterbalanced for each trial session and type of training, so that trainers varied what they taught from session to session. Each trainer conducted five or six training sessions in either vocabulary or mathematics. To further standardize testing and treatment conditions across sessions, all training and other instructions were read from a well-rehearsed script.

## 2.4 Posttest

A 50-item posttest contained 30 vocabulary items (15 items presented in training and 15 new items) and 20 mathematics items (10 presented earlier and 10 new), presenting vocabulary first and mathematics second for all participants in all conditions. This posttest was given to all participants regardless of training. We later found that the correct response for two mathematics items was not listed, however, so those items were removed from analyses.

## 2.5 Procedure

Data collection spanned 22 weeks, with 24 testing sessions having between 7 and 48 people per session. Participants signed up for the experiment between 4 weeks to 1 hour before participating. On arrival, participants completed consent forms and the Demographics Questionnaire I. The consent form included the option to allow researchers to access university records of their high school GPAs, college GPAs, mathematics and English grades, and ACT or SAT college admission scores; 92% of the participants consented. But university records reported ACT scores for only 61.5% of the participants. Having missing data on this variable was not significantly related to the condition to which the participant was later assigned ( $\chi^2 = 1.614$ ,  $p = .204$ ). We substituted self-reported SAT, ACT, and GPAs for those participants who did not consent or who had missing data in university records, and we converted SAT scores to ACT estimated scores using tables provided by ACT and Educational Testing Services (Dorans, Lyu, Pommerich, and Houston 1997). Although it is possible to estimate missing ACT scores using imputation (e.g., Hill et al. 2004), using self-reported ACT scores is transparent and seemed adequate for present purposes. The remaining pretest materials were then distributed.

Although virtually no outcome data were missing, some data on pretreatment covariates were missing for some participants: 130 (62%) of the quasi-experimental participants had complete predictor data, 24 (11%) had missing data on 1 predictor, and 56 (27%) had missing data on more than 1 predictor. But the overall number of missing observations was quite low (2.6% and 3.6% of all covariate measurements of the randomized and quasi-experiments). Therefore, to maintain the focus on the simple comparison of randomized and nonrandomized evaluations, we filled in missing values using EM-based imputation using the missing-data module of SPSS 14.0. These imputations are biased because they do not include an error component. In subsequent research, we intend to examine the sensitivity of propensity score analyses to different ways of treating missing data.

At the end of the time allotted for pretests, participants were assigned randomly to be in a randomized ( $n = 235$ ) or a nonrandomized ( $n = 210$ ) experiment using randomly chosen months of birth; these randomly chosen birth month assignments were counterbalanced over each training session. Participants born in three randomly chosen months were sent to the vocabulary training condition of the randomized experiment ( $n = 116$ ). Participants born in three other randomly chosen months were sent to the mathematics training condition of the randomized experiment ( $n = 119$ ). As they left for the training sessions, these participants were given packets labeled “R” (for randomized experiment) containing posttest materials. Next, the 210 participants who were randomly assigned to the nonrandomized treatment condition were asked to privately select which training session they would prefer to attend and list the reason for their selections. Of these, 131 (62.4%) chose vocabulary training and 79 (37.6%) chose mathematics training. These participants received packets marked “Q” (for quasi-experiment) containing the same posttest materials given to the participants in the randomized experiment, and they were sent to the same training sessions as those who had been randomly assigned to vocabulary or mathematics training. Each training session lasted about 15 minutes. Afterward all participants completed

both the mathematics and vocabulary posttests, submitted them to the trainer, and received debriefing. The trainer marked each posttest as to whether the participant had received mathematics or vocabulary training.

### 3. RESULTS

#### 3.1 Initial Results

Results from the randomized experiment are the presumed best estimate against which all adjusted and unadjusted nonrandomized results are compared. But randomized experiments still encounter group differences in covariates due to sampling error, so we adjusted the randomized results using all of the available covariates in backward stepwise regression. Eventual bias reductions were similar whether we used the adjusted or unadjusted randomized results as a benchmark, however.

*3.1.1 The Effects of Mathematics Training on Mathematics Outcome.* In the covariance-adjusted randomized experiment, participants who received mathematics training performed 4.01 points (out of 18) better on the mathematics outcome than participants who received vocabulary training (Table 1). In the unadjusted nonrandomized experiment, the same effect was 5.01 points, or 25% larger than in the randomized experiment. The

Table 1. Percent bias reduction in quasi-experimental results by propensity score adjustments

	Mean difference (standard error)	Absolute bias ( $\Delta$ )	Percent bias reduction	$R^2$
<b>Mathematics outcome</b>				
Covariate-adjusted randomized experiment	4.01 (.35)	.00		.58
Unadjusted quasi-experiment	5.01 (.55)	1.00		.28
PS stratification	3.72 (.57)	.29	71%	.29
Plus covariates	3.74 (.42)	.27	73%	.66
PS linear ANCOVA	3.64 (.46)	.37	63%	.34
Plus covariates	3.65 (.42)	.36	64%	.64
PS nonlinear ANCOVA	3.60 (.44)	.41	59%	.34
Plus covariates	3.67 (.42)	.34	66%	.63
PS weighting	3.67 (.71)	.34	66%	.16
Plus covariates	3.71 (.40)	.30	70%	.66
PS stratification with predictors of convenience	4.84 (.51)	.83	17%	.28
Plus covariates	5.06 (.51)	1.05	-5%*	.35
ANCOVA using observed covariates	3.85 (.44)	.16	84%	.63
<b>Vocabulary Outcome</b>				
Covariate-adjusted randomized experiment	8.25 (.37)			.71
Unadjusted quasi-experiment	9.00 (.51)	.75		.60
PS stratification	8.15 (.62)	.11	86%	.55
Plus covariates	8.11 (.52)	.15	80%	.76
PS linear ANCOVA	8.07 (.49)	.18	76%	.62
Plus covariates	8.07 (.47)	.18	76%	.76
PS nonlinear ANCOVA	8.03 (.50)	.21	72%	.63
Plus covariates	8.03 (.48)	.22	70%	.77
PS weighting	8.22 (.66)	.03	96%	.54
Plus covariates	8.19 (.51)	.07	91%	.76
PS stratification with predictors of convenience	8.77 (.48)	.52	30%	.62
Plus covariates	8.68 (.47)	.43	43%	.65
ANCOVA using observed covariates	8.21 (.43)	.05	94%	.76

NOTE: All estimates are based on regression analyses. For propensity score stratification stratum weights according to propensity score quintiles were used. Standard errors for propensity score methods are based on 1,000 bootstrap samples (separate samples for each group), with refitted propensity scores and quintiles for each sample (predictors remained unchanged). Each model is presented with only the propensity scores used in the adjustment, and then with the same propensity score adjustment plus the addition of covariates based on backward stepwise inclusion (with main effects only).

\*This adjustment increased bias by 5%.

absolute value of the difference between these results ( $\Delta = |4.01 - 5.01| = 1.00$ ) is a measure of the bias in the unadjusted nonrandomized results, where  $\Delta = 0$  indicates no bias.

*3.1.2 The Effects of Vocabulary Training on Vocabulary Outcome.* In the covariance-adjusted randomized experiment, the participants who received vocabulary training performed 8.25 points (out of 30) better on the vocabulary outcome than the participants who received mathematics training (see Table 1). In the nonrandomized experiment, the same effect was 9.00 points, or 9% larger than in the randomized experiment. The absolute value of the difference between these results is  $\Delta = |8.25 - 9.00| = .75$ .

## 3.2 Adjusted Results

There is only borderline evidence indicating that the results from the nonrandomized experiment differ significantly different from those of the randomized experiment. Still, of particular interest in this study is whether the results from the nonrandomized experiment can be made to more closely approximate results from the randomized experiment. We now explore several alternative adjustments to assess the extent to which they offer reductions in the estimated bias.

*3.2.1 Using Ordinary Linear Regression.* Many researchers would adjust the nonrandomized results using ordinary linear regression, predicting outcome from treatment condition and the observed covariates. This method, with backward selection of main effects only, reduced the estimated bias by 94% for vocabulary outcome and 84% for mathematics outcome. In Table 1, this is the best adjustment for mathematics outcome and the second-best adjustment for vocabulary outcome.

*3.2.2 Using Propensity Scores.* Although several other kinds of adjustments are possible, such as econometric selection bias modeling (e.g., Heckman et al. 1997), we focus on propensity score analysis because of the transparency of its methods and assumptions, its current popularity, and the ease with which it can be done. For person  $i$  ( $i = 1, \dots, N$ ), let  $Z_i$  denote the treatment assignment ( $Z_i = 1$  if the person receives treatment, in our study vocabulary training, and  $Z_i = 0$  if the person receives no or another treatment, here mathematics training) and let  $\mathbf{x}_i$  denote the vector of observed covariates. The propensity score for person  $i$  is the conditional probability of receiving the treatment given the vector of observed covariates,  $e(\mathbf{x}_i) = \Pr(Z_i = 1 | \mathbf{X}_i = \mathbf{x}_i)$ , where it is assumed that, given the  $\mathbf{X}$ 's, the  $Z_i$ 's are independent. Various authors (e.g., Rosenbaum and Rubin 1983) have shown that methods that equate groups on  $e(\mathbf{X})$ , like subclassification, weighting, or regression adjustment, tend to produce unbiased estimates of the treatment effects if the assumption of strongly ignorable treatment assignment holds. This is the case if treatment assignment ( $Z$ ) and the potential outcomes [ $Y = (Y_0, Y_1)$ , under the control and treatment condition] are conditionally independent given the observed covariates  $\mathbf{X}$ , that is,  $\Pr(Z | \mathbf{X}, Y) = \Pr(Z | \mathbf{X})$ , and if  $0 < \Pr(e(\mathbf{x}_i)) < 1$ , for all  $\mathbf{x}_i$ . The assumption is met if all variables related to both those outcomes and treatment assignment are included among the covariates (i.e., there is no hidden bias) and if there is a nonzero probability of being assigned to the treatment or comparison group for all persons (Rosenbaum and Rubin 1983).

Using these data, we created propensity scores using logistic regression. All subsequent analyses used logit-transformed propensity scores (Rubin 2001). Correlations between predictors and both choice of condition and outcome variables are given in Table 2. Without looking at the outcome variables, we tried many models for creating propensity scores, selecting the one that maximized balance on Rubin's (2001) criteria: (a) The standardized difference in the mean propensity score in the two groups (B) should be near 0, (b) the ratio of the variance of the propensity score in the two groups (R) should be near 1, and (c) ratio of the variances of the covariates after adjusting for the propensity score must be close to 1, where ratios between .80 and 1.25 are desirable and those  $< .50$  or  $< 2.0$  are far too extreme. The propensity scores that we used were well balanced using these criteria (Table 3), except that three covariates had variance ratios slightly outside the desirable range (extraversion, 1.357; openness to experience, .799; number of prior math courses, 1.324). They also were well balanced using the criteria proposed by Rosenbaum and Rubin (1984); a  $2 \times 5$  analysis of variance (treatment conditions by propensity score quintiles) yielded no significant main effect for treatment and no interaction for any of the covariates in this study. Figure 2 presents a kernel density graph of the propensity score logits both for the total sample (with vertical quintile borders) and by condition. Overlap was reasonable except at the extremes, and quintiles all had at least five units in each cell.

Table 1 reports four propensity score adjustments for the nonrandomized experiment: (a) stratification on propensity score quintiles (Rosenbaum and Rubin 1984), (b) use of the propensity score as a covariate in an analysis of covariance (ANCOVA), (c) propensity score ANCOVA including nonlinear (quadratic and cubic) terms, and (d) propensity score weighting (Rubin 2001). Table 1 reports all four adjustments by themselves, and then all four in a model that also includes some of the original covariates entered in a backward-stepwise manner (the rows labeled "Plus covariates"). The table also reports the usual regression-based standard errors, except that standard errors for methods involving propensity scores were bootstrapped. For each bootstrap sample, the propensity scores were refit; the predictors included remained unchanged.

Overall, the eight propensity score adjustments reduced bias by an average of 74% (range, 59–96%), depending on the model. Bias reduction was higher for vocabulary outcome ( $M = 81\%$ ; range, 70–96%) than for mathematics outcome ( $M = 66\%$ ; range, 59–73%). Differences in the specific adjustment used were minor and probably should be treated as non-significant given the standard errors, although stratification and weighting tended to perform better than ANCOVA. The addition of covariates to any of the propensity score adjustments significantly increased the variance accounted for, made little difference in bias reduction, and slightly reduced the bootstrapped standard errors of the estimate. Standard errors for propensity score weighting were larger than for any other method, likely inflated by the presence of some very low propensity scores. Standard errors also were high for propensity score stratification, reflecting increased uncertainty about the treatment effect given the coarseness of the strata and the small samples in some cells. Otherwise, standard errors for propensity score-adjusted effects were moderately larger than those for the original covariate-adjusted randomized experiments.

Table 2. Correlations between predictors and outcome in nonrandomized experiments

Predictor	Vocabulary posttest	Mathematics posttest	Chose vocabulary training
Vocabulary pretest	.468**	.109	.169*
Mathematics pretest	.147*	.446**	-.090
Number of prior mathematics courses <sup>†</sup>	-.018	.299**	-.131
Like mathematics	-.288**	.471**	-.356**
Like literature	.233**	-.226**	.164*
Preferring literature over mathematics	.419**	-.426**	.385**
Extraversion	.005	-.158*	.092
Agreeableness	.120	-.078	.098
Conscientiousness	-.189**	-.041	-.126
Emotionality	-.099	-.115	-.015
Openness to experience	.201**	.050	.053
Mathematics anxiety	-.051	-.140*	.003
Depression <sup>†</sup>	.087	.149*	-.014
Caucasian	.322**	-.074	.178*
African-American	-.296**	-.015	-.144*
Age <sup>†</sup>	.077	-.217**	.022
Male	.064	.141*	-.065
Married	-.073	-.162*	.001
Mother education	.094	-.022	.010
Father education	.110	.068	.008
College credit hours <sup>†</sup>	.132	.125	.033
Math-intensive major	-.169*	.298**	-.191**
ACT comprehensive score	.341**	.418**	.028
High school GPA	-.003	.401**	-.041
College GPA	.059	.219**	-.026

\*  $P < .05$ ; \*\*  $P < .01$  (two-tailed).

<sup>†</sup>These four variables were log-transformed in all analyses to reduce positive skew.

Selecting covariates to use in creating propensity scores is a crucial aspect of good propensity score analysis (Brookhart et al. 2006). The present study was designed to have a rich set of covariates potentially related to treatment choice and outcome. Yet in practice, many researchers create propensity scores from whatever variables are conveniently available. To explore the potential consequences of using only conveniently available covariates, we created a new set of propensity scores using only sex, age, marital status, and race (dummy coded for two predictors, Caucasian and African-American) as predictors. Those

variables often are gathered in research and are the kinds of predictors of convenience likely to be available when careful thought has not gone into the inclusion of potential selection variables. Adjusting the results of the nonrandomized experiment by stratifying according to the quintiles of such propensity scores yielded inconsistent, and usually poor, results (Table 1). For the mathematics outcome, this adjustment reduced bias by 17% (and increased bias by 5% when covariates were added); for the vocabulary outcome, this adjustment reduced bias by 30% (43% when covariates were added). Some bias reduction

Table 3. Rubin's (2001) balance criteria before and after propensity score stratification

Analysis	Propensity score		Number of covariates with variance ratio				
	<i>B</i>	<i>R</i>	$\leq 1/2$	$> 1/2$ and $\leq 4/5$	$> 4/5$ and $\leq 5/4$	$> 5/4$ and $\leq 2$	$> 2$
Before any adjustment	-1.13	1.51	0	2	17	6	0
After stratification on propensity scores constructed from all covariates	-.03	.93	0	1	22	2	0
After stratification on propensity scores constructed from predictors of convenience, balance tested only on the five predictors of convenience	-.01	1.10	0	0	5	0	0
After stratification on propensity scores constructed from predictors of convenience, balance tested on all 25 covariates	-.01	1.10	0	2	16	7	0

NOTE: Standardized mean difference in propensity scores are given by  $B = (\bar{x}_t - \bar{x}_c) / \sqrt{(s_t^2 + s_c^2)/2}$  where  $\bar{x}_t$  and  $\bar{x}_c$  are the sample means of the propensity scores in the treatment and comparison group, and  $s_t^2$  and  $s_c^2$  the corresponding sample variances. The variance ratio, *R*, is  $s_t^2/s_c^2$  (also for covariates). Balancing criteria after propensity score stratification are obtained by attaching stratum weights to individual observations (Rubin 2001).

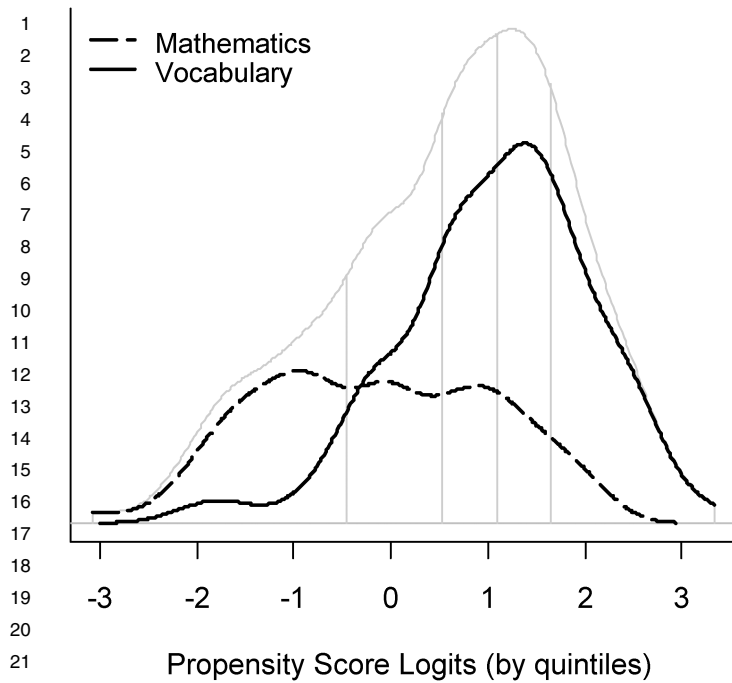


Figure 2. Distribution of propensity score logits smoothed using a kernel density function. The light-gray line represents the total sample, with vertical quintile borders. The dashed line represents those who chose mathematics training, and the solid black line represents those who chose vocabulary training. Negative scores indicate propensity to choose mathematics training.

occurred because these four predictors are related to selection (Table 2), but those four predictors clearly are not the only relevant ones.

If a researcher had tested the propensity scores resulting from the five predictors of convenience using Rubin's (2001) balance criteria, they would have performed quite well (Table 3, third row). But this would have hidden a failure to balance well on many of the remaining covariates that presumably would not have been observed by such a researcher (Table 3, fourth row). This is a good illustration of hidden bias and how it might lead to poor estimates of a treatment effect.

## 4. DISCUSSION

### 4.1 Adjustments to Nonrandomized Experiments

This study suggests that adjusted results from nonrandomized experiments can approximate results from randomized experiments. This was true for propensity score adjustments, as well as for ordinary linear regression without the use of propensity scores, some implications of which we discuss shortly. All of the recommended adjustments always reduced bias (and never increased it), and did so substantially. Moreover, they did so despite the fact that the nonrandomized study had a small sample size and was not designed to have a well-matched control group before data collection began. These adjustments might have been even better had the study been designed to be larger with a well-matched control group.

The adjustments may have done well in the present case in part because this study is characterized by a very rich set of covariates that are well measured and plausibly related to both the

selection process and the outcome measures. Such richness is not always present in data sets for nonrandomized experiments, especially not in those conducted retrospectively. As demonstrated by our analysis of propensity scores based on predictors of convenience, a lack of covariate richness may greatly reduce the accuracy of adjustments. Implicit is a lesson for the prospective design of nonrandomized experiments, that attention to careful measurement of the selection process can be crucial to the success of subsequent analyses.

Furthermore, our experience analyzing this data set suggests that propensity score adjustments may be sensitive to variations in how those scores are constructed. One example is the sensitivity to which covariate balance criteria are used. We found that some propensity scores constructed under Rosenbaum and Rubin's (1984) balance criteria did not meet Rubin's (2001) balance criteria, but those meeting the latter criteria always met the former. The reliance of the criteria of Rosenbaum and Rubin (1984) on significance testing makes it vulnerable to confusing successful balance with low power. The emphasis of Rubin (2001) on the size of the imbalance may be more desirable. Both sets of criteria probably should be reported. We would benefit from further development of ways to create and assess balance (e.g., Imai, King, and Stuart 2007; Sekhon 2007), as well as from better-justified standards for how much balance should be achieved.

The results also were sensitive to how missing data in the predictors were managed. At first, we followed a recommendation of Rosenbaum and Rubin (1984) to create propensity scores separately for groups with different missing-data patterns. But we found that bias reduction was highly sensitive to seemingly minor changes in how those patterns were identified, in one case even increasing bias. Consequently, we moved to more current missing-data methods, but those results also may prove sensitive to which current method is used (D'Agostino and Rubin 2000). In particular, our results might have changed had we used multiple imputation rather than EM-based imputation.

We used logistic regression to construct propensity scores in the present study. Other methods for creating propensity scores exist, including classification trees, boosted regression, random forests, and boosted regression (e.g., Stone et al. 1995; McCaffrey, Ridgeway, and Morral 2004). A simulation conducted by one of our colleagues suggests that propensity score adjustments also may be sensitive to the methods used, and also quite sensitive to sample size (Luellen 2007).

We are currently exploring the sensitivity of the present data set to many of the variations described in the previous paragraphs. Taking them together, however, it may be that the practice of propensity score analysis in applied research may yield adjustments of unknown or highly variable accuracy. This is not surprising for a method as new as propensity score analysis, and points to the need for more clarity about best propensity score practice.

In view of these matters, a pertinent question is why researchers should consider using propensity scores when ordinary linear regression with covariates does as well or better. One situation in which propensity scores could be used is when the design calls for matching treatment and comparison units on a large number of covariates, for example, when constructing a

60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118

control group matched to an existing treatment group from a large number of potential controls (e.g., Rubin 2001). Without reducing those covariates to a propensity score, the matching process would not be feasible. Another circumstance is when there is uncertainty about assumptions of linearity in ordinary linear regression that stratification on propensity scores might ameliorate. Such exceptions aside, however, in general our results do not support the preferential use of propensity scores over ordinary linear regression.

#### 4.2 Comments on the Laboratory Analog Design Used in This Study

Questions may arise about the replicability and generalizability of these results given the design that we used. The design probably is no more labor-intensive than other methods, at least for researchers with access to large research participant pools like those available in university-based introductory psychology classes. Thus testing replication has few obstacles. Minor changes in the method might improve its feasibility and yield. The second author, for example, added a no-treatment control group to this design in a study in progress and also added achievement motivation as an additional predictor of selection. The first author is working to computerize administration of this method, which might allow rapid implementation of more complex assignment mechanisms or allow Web-based implementation to obtain larger sample sizes. We are also creating a version of the study that can be administered over the Internet, allowing us to improve certain features of this study; for example, we can use computer-generated random numbers rather than birth month to do random assignment.

The question of generalization is more serious and has two aspects. The first aspect concerns how the results reported in this study would change over variations of the method that stay within this general laboratory analog paradigm. One could vary the kind of treatment from the current educational one to mimic other substantive areas, such as job training, health, and different aspects of education. Similarly, one could create more time-consuming treatments, although it would be desirable to avoid attrition from both treatment and measurement, because these are separate problems from adjusting for selection into conditions.

A second variation within the laboratory analog method is to study different selection mechanisms, such as cutoff-based assignment mechanisms used in the regression discontinuity design (Shadish et al. 2002), analogs to parental selection of children into interventions, and analogs to the kind of selection that occurs in mental health, where participants choose treatment due to extremely high scores on a latent variable such as distress. Such work could advance an empirical theory of selection for different kinds of treatments, improving the efficacy of adjustments that rely on good prediction of selection.

A third variation of the present method is to explore different design elements or analyses. For example, propensity score matching may benefit when the researcher has a much larger pool of potential control group participants from which to select propensity score matches to a smaller group of treatment group participant scores (Rosenbaum and Rubin 1985; Rubin 2001). This should be easy to test with a variation of the present method. Given that propensity score adjustments also are said

to work best in large samples, one could also vary sample size to shed light on sample size requirements and randomly assign proportionately more participants to the nonrandomized experiment. The latter also would decrease the standard errors of adjusted estimates. Similarly, one might examine the effectiveness of additional statistical adjustment procedures, such as econometric selection bias models (e.g., Heckman et al. 1997; Greene 1999).

A fourth variation of our method is to study populations other than introductory psychology students. We used psychology students because we could obtain large numbers of them and could exercise a high degree of experimental control. Other populations can approximate those characteristics, especially if the treatment is short or participation is required. For example, Akins, Hollandsworth, and O'Connell (1982) treated introductory psychology and sociology students solicited for dental fear with a 1-hour, researcher-administered intervention given by audio and videotape in a college laboratory. This could be offered to university or community participants more generally. Aiken, West, Schwalm, Carroll, and Hsiung (1998) used students who were required to take a university remedial writing program to create a study similar to the present one, but without the initial random assignment to assignment method. Such cases may be adapted to remedy the latter lacuna. So might the provision of desirable brief services to community participants, such as stress reduction training, especially if accompanied by payment for participation. One could argue that such examples are not really laboratory analogs anymore—especially if they were also conducted in the community rather than in the laboratory—but if so, so much the better.

The latter observation leads into the second part of the generalization question—whether highly controlled laboratory experiments like the present study yield results that would replicate in research about the effects of longer treatments in settings like the classroom, job training center, or physician office where field experimentation takes place. Some variations on our basic laboratory analog could shed light on this concern, such as the hypothetical medical experiment described in Section 1 at the end of the discussion of doubly randomized preference trials. But attrition from measurement and treatment are prevalent in such applied settings and add additional layers of selection bias that propensity scores were not necessarily designed to adjust, as noted for the study of Janevic et al. (2003) (see also Long et al., in press). Ultimately, the only way to answer this generalization question is to apply the paradigm in the present study to actual field experiments. Such a study might be hard to sell to funding agencies, especially to problem-focused agencies that might be reluctant to spend extra money to fund the nonrandomized experiment if they are already funding the randomized one. Nonetheless, we suspect that chances to do such studies will present themselves in due course to researchers who are sensitive to the opportunity.

[Received July 2007. Revised December 2007.]

#### REFERENCES

- Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., and Hsiung, S. (1998). "Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation: Efficacy of a University-Level Remedial Writing Program." *Evaluation Review*, 22, 207–244.

- 1 Akins, T., Hollandsworth, J. G., and O'Connell, S. J. (1982), "Visual and Verbal  
2 Modes of Information Processing and Their Relation to the Effectiveness of  
3 Cognitively-Based Anxiety-Reduction Techniques," *Behaviour Research and  
4 Therapy*, 20, 261–268.
- 5 Beck, A. T., and Beck, R. W. (1972), "Screening Depressed Patients in Family  
6 Practice: A Rapid Technique," *Postgraduate Medicine*, 51, 81–85.
- 7 Bloom, H. S., Michalopoulos, C., Hill, C. J., and Lei, Y. (2002), *Can Nonexper-  
8 imental Comparison Group Methods Match the Findings From a Random  
9 Assignment Evaluation of Mandatory Welfare-to-Work Programs?*, New York:  
10 Manpower Development Research Corp.
- 11 Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J.,  
12 and Stürmer, T. (2006), "Variable Selection for Propensity Score Models,"  
13 *American Journal of Epidemiology*, 163, 1149–1156.
- 14 D'Agostino, R. B., and Rubin, D. B. (2000), "Estimating and Using Propen-  
15 sity Scores With Partially Missing Data," *Journal of the American Statistical  
16 Association*, 95, 749–759.
- 17 Dehejia, R., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies:  
18 Reevaluating the Evaluation of Training Programs," *Journal of the American  
19 Statistical Association*, 94, 1053–1062.
- 20 Dorans, N. J., Lyu, C. F., Pommerich, M., and Houston, W. M. (1997), "Concor-  
21 dance Between ACT Assessment and Recentered SAT I Sum Scores," *College  
22 and University*, 73, 24–33.
- 23 Drake, ?? (1993), ???.
- 24 Educational Testing Service (1962), "Vocabulary Test II (V-2)," in *Kit of Factor  
25 Referenced Cognitive Tests*, Princeton, NJ: Author.
- 26 ——— (1993), "Arithmetic Aptitude Test (RG-1)," in *Kit of Factor Referenced  
27 Cognitive Tests*, Princeton, NJ: Author.
- 28 Faust, M. W., Ashcraft, M. H., and Fleck, D. E. (1996), "Mathematics Anxiety  
29 Effects in Simple and Complex Addition," *Mathematical Cognition*, 2, 25–62.
- 30 Glazer, S., Levy, D. M., and Myers, D. (2003), "Nonexperimental versus  
31 Experimental Estimates of Earnings Impacts," *The Annals of the American  
32 Academy of Political and Social Science*, 589, 63–93.
- 33 Goldberg, L. R. (1997), "Big-Five Factor Markers Derived From the IPIP Item  
34 Pool (Short Scales)," *International Personality Item Pool: A Scientific Collab-  
35 oratory for the Development of Advanced Measures of Personality and Other  
36 Individual Differences*, available at [http://ipip.ori.org/ipip/appendixa.htm#  
37 AppendixA](http://ipip.ori.org/ipip/appendixa.htm#AppendixA).
- 38 Greene, W. H. (1999), *Econometric Analysis*, Upper Saddle River, NJ: Prentice-  
39 Hall.
- 40 Heckman, J. J., Ichimura, H., and Todd, P. E. (1997), "Matching as an Econo-  
41 metric Evaluation Estimator: Evidence From Evaluating a Job Training Pro-  
42 gramme," *Review of Economic Studies*, 64, 605–654.
- 43 Heinsman, D. T., and Shadish, W. R. (1996), "Assignment Methods in Experi-  
44 mentation: When Do Nonrandomized Experiments Approximate the Answers  
45 From Randomized Experiments?" *Psychological Methods*, 1, 154–169.
- 46 Hill, J. L., Reiter, J. P., and Zanutto, E. L. (2004), "A Comparison of Experi-  
47 mental and Observational Data Analyses," in *Applied Bayesian Modeling  
48 and Causal Inference From Incomplete Data Perspectives*, eds. A. Gelman  
49 and X.-L. Meng, New York: Wiley, pp. 51–60.
- 50 Hill, J. L., Rubin, D. B., and Thomas, N. (2000), "The Design of the New York  
51 School Choice Scholarship Program Evaluation," in *Validity and Social Ex-  
52 perimentation: Donald Campbell's Legacy*, Vol. 1, ed. L. Bickman, Thousand  
53 Oaks, CA: Sage, pp. 155–180.
- 54 Hirano, K., and Imbens, G. W. (2001), "Estimation of Causal Effects Using  
55 Propensity Score Weighting: An Application to Data on Right Heart Catheter-  
56 ization," *Health Services & Outcomes Research Methodology*, 2, 259–278.
- 57 Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the Ameri-  
58 can Statistical Association*, 81, 945–970.
- 59 Imai, K., King, G., and Stuart, E. A. (2007), "Misunderstandings Among Ex-  
60 perimentalists and Observationalists About Causal Inference," ???, available  
61 at <http://imai.princeton.edu/research/files/matchse.pdf>.
- 62 Janevic, M. R., Janz, N. K., Lin, X., Pan, W., Sinco, B. R., and Clark, N. M.  
63 (2003), "The Role of Choice in Health Education Intervention Trials: A Re-  
64 view and Case Study," *Social Science and Medicine*, 56, 1581–1594.
- 65 Kleinmuntz, B. (1960), "Identification of Maladjusted College Students," *Jour-  
66 nal of Counseling Psychology*, 7, 209–211.
- 67 Lipsey, M. W., and Wilson, D. B. (1993), "The Efficacy of Psychological,  
68 Educational, and Behavioral Treatment: Confirmation From Meta-Analysis,"  
69 *American Psychologist*, 48, 1181–1209.
- 70 Long, Q., Little, R. J., and Lin, X. (in press), "Causal Inference in Hybrid In-  
71 tervention Trials Involving Treatment Choice," *Journal of the American Sta-  
72 tistical Association*, ??, ??–??.
- 73 Luellen, J. K. (2007), "A Comparison of Propensity Score Estimation and  
74 Adjustment Methods on Simulated Data," unpublished doctoral dissertation,  
75 University of Memphis, ???.
- 76 McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity Score  
77 Estimation With Boosted Regression for Evaluating Causal Effects in Obser-  
78 vational Studies," *Psychological Methods*, 9, 403–425.
- 79 Rosenbaum, P. R. (2002), *Observational Studies* (2nd ed.), New York: Springer-  
80 Verlag.
- 81 Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity  
82 Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.
- 83 ——— (1984), "Reducing Bias in Observational Studies Using Subclassifica-  
84 tion on the Propensity Score," *Journal of the American Statistical Association*,  
85 79, 516–524.
- 86 ——— (1985), "The Bias Due to Incomplete Matching," *Biometrics*, 41, 103–  
87 116.
- 88 Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized  
89 and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–  
90 701.
- 91 ——— (2001), "Using Propensity Scores to Help Design Observational Stud-  
92 ies: Application to the Tobacco Litigation," *Health Services and Outcomes  
93 Research Methodology*, 2, 169–188.
- 94 Rucker, G. (1989), "A Two-Stage Trial Design for Testing Treatment, Self-  
95 Selection and Treatment Preference Effects," *Statistics in Medicine*, 8, 477–  
96 485.
- 97 Sekhon, J. S. (2007), "Alternative Balance Metrics for Bias Reduction in  
98 Matching Methods for Causal Inference," ???, available at [http://sekhon.  
99 berkeley.edu/papers/SekhonBalanceMetrics.pdf](http://sekhon.berkeley.edu/papers/SekhonBalanceMetrics.pdf).
- 100 Shadish, W. R. (2000), "The Empirical Program of Quasi-Experimentation,"  
101 in *Validity and Social Experimentation: Donald Campbell's Legacy*, ed. L.  
102 Bickman, Thousand Oaks, CA: Sage, pp. 13–35.
- 103 Shadish, W. R., and Ragsdale, K. (1996), "Random versus Nonrandom Assign-  
104 ment in Controlled Experiments: Do You Get the Same Answer?" *Journal of  
105 Consulting and Clinical Psychology*, 64, 1290–1305.
- 106 Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002), *Experimental  
107 and Quasi-Experimental Designs for Generalized Causal Inference*, Boston:  
108 Houghton-Mifflin.
- 109 Stone, R. A., Obrosky, D. S., Singer, D. E., Kapoor, W. N., Fine, M. J., and the  
110 Pneumonia Patient Outcomes Research Team (PORT) Investigators (1995),  
111 "Propensity Score Adjustment for Pretreatment Differences Between Hos-  
112 pitalized and Ambulatory Patients With Community-Acquired Pneumonia,"  
113 *Medical Care*, 33, AS56–AS66.
- 114 Wennberg, J. E., Barry, M. J., Fowler, F. J., and Mulley, A. (1993), "Outcomes  
115 Research, PORTs, and Health Care Reform," *Annals of the New York Acad-  
116 emy of Sciences*, 703, 52–62.